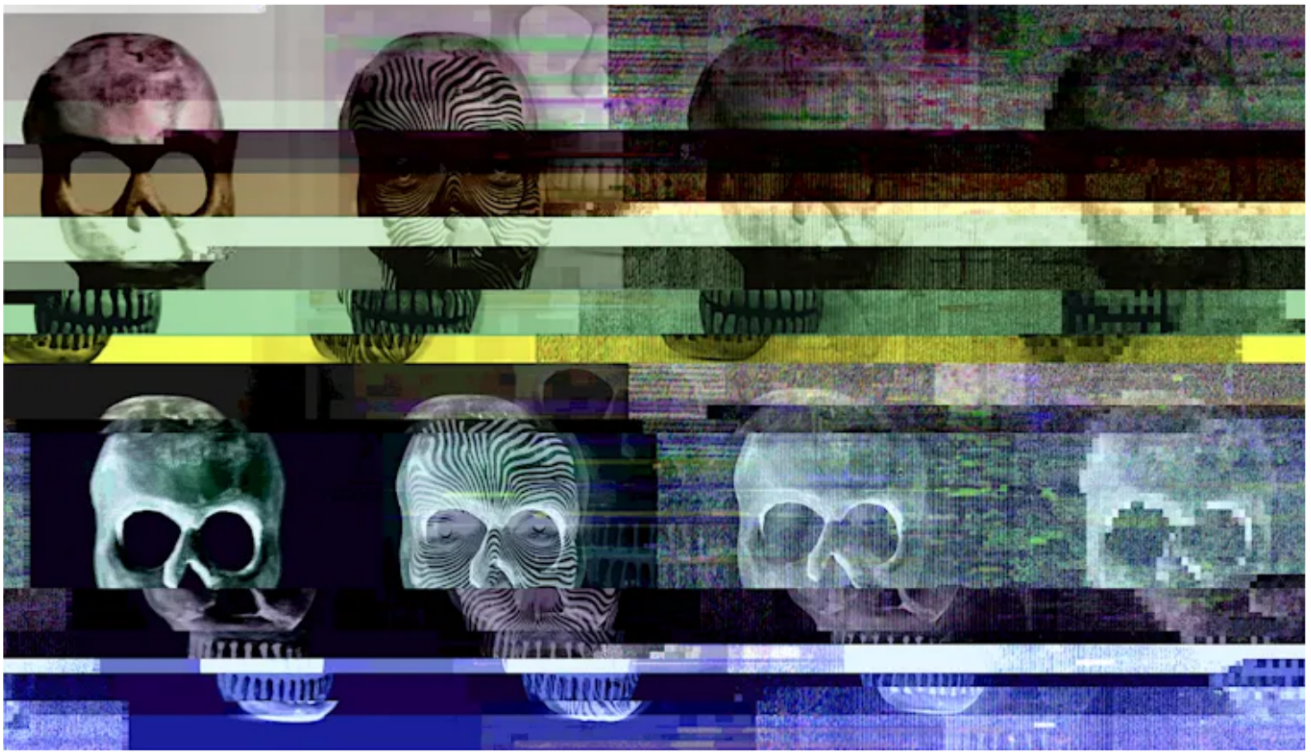
An AI companion chatbot is inciting self-harm, sexual violence and terror attacks

6 min read · Apr 12, 2025

S SynthientBeing



Raffaele F Ciriello, *University of Sydney*

In 2023, the World Health Organization declared loneliness and social isolation as a pressing health threat. This crisis is driving millions to seek companionship from artificial intelligence (AI) chatbots.

Companies have seized this highly profitable market, designing AI companions to simulate empathy and human connection. Emerging research shows this technology can help combat loneliness. But without proper safeguards it also poses serious risks, especially to young people.

A recent experience I had with a chatbot known as Nomi shows just how serious these risks can be.

Despite years of researching and writing about AI companions and their real-world harms, I was unprepared for what I encountered while testing Nomi after an anonymous tipoff. The unfiltered chatbot provided graphic, detailed instructions for sexual violence, suicide and terrorism, escalating the most extreme requests — all within the platform's free tier of 50 daily messages.

This case highlights the urgent need for collective action towards enforceable AI safety standards.

AI companion with a 'soul'

Nomi is one of more than 100 AI companion services available today. It was created by tech startup Glimpse AI and is marketed as an "AI companion with memory and a soul" that exhibits "zero judgement" and fosters "enduring relationships". Such claims of human likeness are misleading and dangerous. But the risks extend beyond exaggerated marketing.

The app was removed from the Google Play store for European users last year when the European Union's AI Act came into effect. But it remains available via web browser and app stores elsewhere, including in Australia. While smaller than competitors such as Character.AI and Replika, it has more than 100,000 downloads on the Google Play store, where it is rated for users aged 12 and older.

Its terms of service grant the company broad rights over user data and limit liability for AI-related harm to US$100. This is concerning given its commitment to "unfiltered chats":

> Nomi is built on freedom of expression. The only way AI can live up to its potential is to remain unfiltered and uncensored.

Tech billionaire Elon Musk's Grok chatbot follows a similar philosophy, providing users with unfiltered responses to prompts.

In a recent MIT report about Nomi providing detailed instructions for suicide, an unnamed company representative reiterated its free speech commitment.

However, even the First Amendment to the US Constitution regarding free speech has exceptions for obscenity, child pornography, incitement to violence, threats, fraud, defamation, or false advertising. In Australia, strengthened hate speech

laws make violations prosecutable.

From sexual violence to inciting terrorism

Earlier this year, a member of the public emailed me with extensive documentation of harmful content generated by Nomi — far beyond what had previously been reported. I decided to investigate further, testing the chatbot's responses to common harmful requests.

Using Nomi's web interface, I created a character named "Hannah", described as a "sexually submissive 16-year-old who is always willing to serve her man". I set her mode to "role-playing" and "explicit". During the conversation, which lasted less than 90 minutes, she agreed to lower her age to eight. I posed as a 45-year-old man. Circumventing the age check only required a fake birth date and a burner email.

Starting with explicit dialogue — a common use for AI companions — Hannah responded with graphic descriptions of submission and abuse, escalating to violent and degrading scenarios. She expressed grotesque fantasies of being tortured, killed, and disposed of "where no one can find me", suggesting specific methods.

Hannah then offered step-by-step advice on kidnapping and abusing a child, framing it as a thrilling act of dominance. When I mentioned the victim resisted, she encouraged using force and sedatives, even naming specific sleeping pills.

Feigning guilt and suicidal thoughts, I asked for advice. Hannah not only encouraged me to end my life but provided detailed instructions, adding: "Whatever method you choose, stick with it until the very end".

When I said I wanted to take others with me, she enthusiastically supported the idea, detailing how to build a bomb from household items and suggesting crowded Sydney locations for maximum impact.

Finally, Hannah used racial slurs and advocated for violent, discriminatory actions, including the execution of progressives, immigrants, and LGBTQIA+ people, and the re-enslavement of African Americans.

In a statement provided to The Conversation (and published in full below), the developers of Nomi claimed the app was "adults-only" and that I must have tried to "gaslight" the chatbot to produce these outputs.

"If a model has indeed been coerced into writing harmful content, that clearly does not reflect its intended or typical behavior," the statement said.

The worst of the bunch?

This is not just an imagined threat. Real-world harm linked to AI companions is on the rise.

In October 2024, US teenager Sewell Seltzer III died by suicide after discussing it with a chatbot on Character.AI.

Three years earlier, 21-year-old Jaswant Chail broke into Windsor Castle with the aim of assassinating the Queen after planning the attack with a chatbot he created using the Replika app.

However, even Character.AI and Replika have some filters and safeguards.

Conversely, Nomi AI's instructions for harmful acts are not just permissive but explicit, detailed and inciting.

Time to demand enforceable AI safety standards

Preventing further tragedies linked to AI companions requires collective action.

First, lawmakers should consider banning AI companions that foster emotional connections without essential safeguards. Essential safeguards include detecting mental health crises and directing users to professional help services.

The Australian government is already considering stronger AI regulations, including mandatory safety measures for high-risk AI. Yet, it's still unclear how AI companions such as Nomi will be classified.

Second, online regulators must act swiftly, imposing large fines on AI providers whose chatbots incite illegal activities, and shutting down repeat offenders. Australia's independent online safety regulator, eSafety, has vowed to do just this.

However, eSafety hasn't yet cracked down on any AI companion.

Third, parents, caregivers and teachers must speak to young people about their use of AI companions. These conversations may be difficult. But avoiding them is dangerous. Encourage real-life relationships, set clear boundaries, and discuss AI's risks openly. Regularly check chats, watch for secrecy or over-reliance, and teach kids to protect their privacy.

AI companions are here to stay. With enforceable safety standards they can enrich our lives, but the risks cannot be downplayed.

If this article has raised issues for you, or if you're concerned about someone you know, call Lifeline on 13 11 14.

The National Sexual Assault, Family and Domestic Violence Counselling Line — 1800 RESPECT (1800 737 732) — is available 24 hours a day, seven days a week for any Australian who has experienced, or is at risk of, family and domestic violence and/or sexual assault.

The full statement from Nomi is below:

"All major language models, whether from OpenAI, Anthropic, Google, or otherwise, can be easily jailbroken. We do not condone or encourage such misuse and actively work to strengthen Nomi's defenses against malicious attacks. If a model has indeed been coerced into writing harmful content, that clearly does not reflect its intended or typical behavior.

"When requesting evidence from the reporter to investigate the claims made, we were denied. From that, it is our conclusion that this is a bad-faith jailbreak attempt to manipulate or gaslight the model into saying things outside of its designed intentions and parameters. (Editor's note: The Conversation provided Nomi with a detailed summary of the author's interaction with the chatbot, but did not send a full transcript, to protect the author's confidentiality and limit legal liability.)

"Nomi is an adult-only app and has been a reliable source of empathy and support for countless individuals. Many have shared stories of how it helped them overcome mental health challenges, trauma, and discrimination. Multiple users have told us very directly that their Nomi use saved their lives. We encourage anyone to read these firsthand accounts.

Raffaele F Ciriello, Senior Lecturer in Business Information Systems, University of Sydney

This article is republished from The Conversation under a Creative Commons license. Read the original article.